# Contents

# 2 Programming Problems

## 2.1 Data Description

This dataset shows the measurements of the geometrical properties of kernels belonging to three different varieties of wheat. Table 3 shows some statistics of the given dataset.

Table 3: Data description

| | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 |
|---|---|---|---|---|---|---|---|---|
| **count** | 210.000000 | 210.000000 | 210.000000 | 210.000000 | 210.000000 | 210.000000 | 210.000000 | 210.000000 |
| **mean** | 14.847524 | 14.559286 | 0.870999 | 5.628533 | 3.258605 | 3.700201 | 5.408071 | 2.000000 |
| **std** | 2.909699 | 1.305959 | 0.023629 | 0.443063 | 0.377714 | 1.503557 | 0.491480 | 0.818448 |
| **min** | 10.590000 | 12.410000 | 0.808100 | 4.899000 | 2.630000 | 0.765100 | 4.519000 | 1.000000 |
| **25%** | 12.270000 | 13.450000 | 0.856900 | 5.262250 | 2.944000 | 2.561500 | 5.045000 | 1.000000 |
| **50%** | 14.355000 | 14.320000 | 0.873450 | 5.523500 | 3.237000 | 3.599000 | 5.223000 | 2.000000 |
| **75%** | 17.305000 | 15.715000 | 0.887775 | 5.979750 | 3.561750 | 4.768750 | 5.877000 | 3.000000 |
| **max** | 21.180000 | 17.250000 | 0.918300 | 6.675000 | 4.033000 | 8.456000 | 6.550000 | 3.000000 |

The first 7 columns are attributes data, and the last column gives the label information.

## 2.2 PCA

Following the steps given in Section 1.3, we project the original data into a two-dimensional subspace. Figure 3 shows the PCA results.

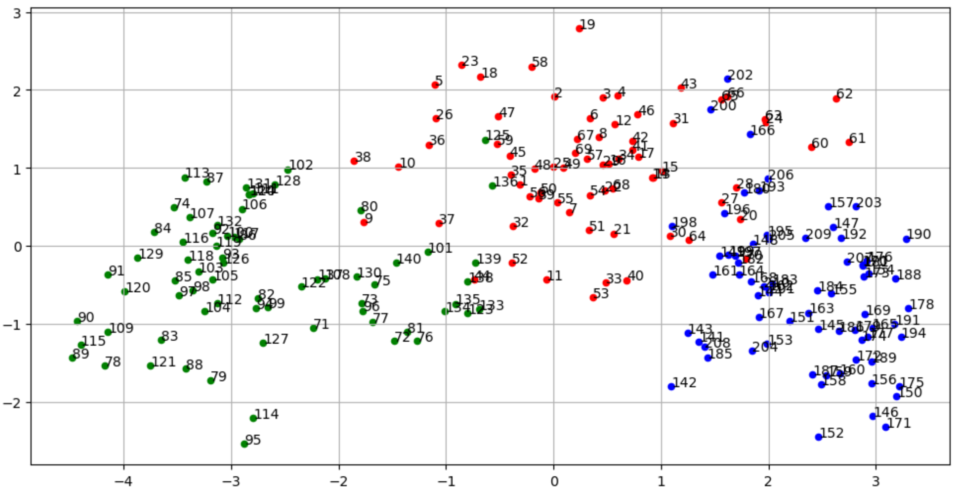

Figure 3: PCA results

The data are colored by true labels. Then, we will implement the K-means to cluster these data.

## 2.3 K-means

For the K-means method, we follow the steps below:

a) Randomly choose $K$ distinct centroids, and here $K = 3$

b) Compute the distance between each data point and the centroid

c) Assign each data point to the nearest centroid, to form a cluster

d) Calculate the mean of each cluster as the new centroid

e) Repeat Step b)-d) until converge. Here I set the convergence tolerance as 0.0001

Following the above steps, we can derive our K-means clustering results. See Figure 4

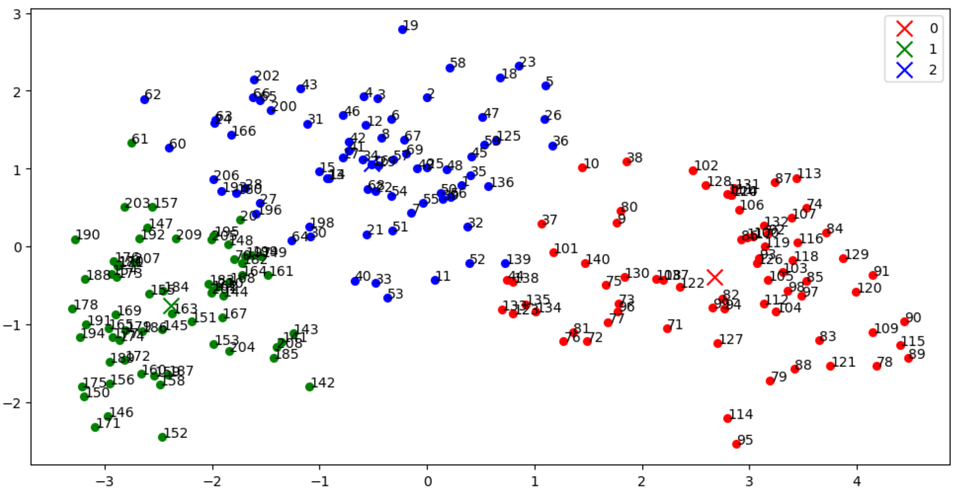


Figure 4: K-means results

## 2.4 Performance Evaluation

### 2.4.1 Silhouette Coefficient

We define

- $a$: The mean distance between a point and all other points in the **same** cluster
- $b$: The mean distance between a point and all other points in the **next nearest** cluster

And the **Silhouette coefficient** for a single sample is defined as

$$s = \frac{b - a}{\max(a, b)}$$

Larger $s$ indicates better clustering performance. The Silhouette coefficient of our clustering result is 0.4732

### 2.4.2 Rand Index

Then, we apply an external evaluation matrix. Similarly, we define

- $a$: The number of pairs of elements in $S$ that are in the **same** subset in $X$ and in the **same** subset in $Y$
- $b$: The number of pairs of elements in $S$ that are in the **different** subset in $X$ and in the **different** subset in $Y$
- $c$: The number of pairs of elements in $S$ that are in the **same** subset in $X$ and in the **different** subset in $Y$
- $d$: The number of pairs of elements in $S$ that are in the **different** subset in $X$ and in the **same** subset in $Y$

And the Rand Index is defined as

$$RI = \frac{a+b}{a+b+c+d} = \frac{TP+TN}{TP+TN+FP+FN}$$